ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Bounds for Rademacher Processes via Chaining

Technical Report

Johannes Christof Lederer
lederer@stat.math.ethz.ch

Abstract: We study Rademacher processes where the coefficients are functions evaluated at fixed, but arbitrary covariables. Specifically, we assume the function class under consideration to be parametrized by the standard cocube in l dimensions and we are mainly interested in the high-dimensional, asymptotic situation, that is, l as well the number of Rademacher variables n go to infinity with l much larger than n. We refine and apply classical entropy bounds and Majorizing Measures, both going back to the well known idea of chaining. That way, we derive general upper bounds for Rademacher processes. In the linear case and under high correlations, we further improve on these bounds. In particular, we give bounds independent of l for highly correlated covariables.

October 2010

1 Introduction

We study upper bounds for the quantity

$$\mathbb{E} \sup_{\theta \in \Theta} \left| \sum_{i=1}^{n} \epsilon_i \phi_\theta(x_i) \right| \tag{1}$$

with $\Theta := \{\theta \in \mathbb{R}^l : \|\theta\|_1 \leq M\}$, i.i.d. Rademacher variables ϵ_i and real valued functions ϕ_θ evaluated at fixed but arbitrary x_i. We are mainly interested in the high-dimensional, asymptotic situation, i.e., $l \gg n$ and $n, l \to \infty$ and we treat a general setting, the linear case as well as a setting involving strongly correlated x_i. We show in particular that strong correlations can lead to better asymptotic bounds.

Chaining is the main tool for our investigations. For an arbitrary process $\{Z_\lambda : \lambda \in \Lambda\}$ it means the following: instead of studying terms of the form $|Z_\lambda - Z_{\lambda'}|$ for (possibly very distinct) random variables $Z_\lambda, Z_{\lambda'}$ directly, one applies the triangular inequality

$$|Z_\lambda - Z_{\lambda'}| \leq \sum_{j=1}^{m} |Z_{\lambda_n} - Z_{\lambda_{n-1}}|$$

and studies the increments $|Z_{\lambda_n} - Z_{\lambda_{n-1}}|$, where $\lambda_n, \lambda_{n-1} \in \Lambda$, $\lambda_0 = \lambda$ and $\lambda_m = \lambda'$. Usually, the $Z_{\lambda_0}, ..., Z_{\lambda_m}$ are constructed such that $Z_\lambda - Z_{\lambda'}$ can be thought of as the sum of the small "chain links" $Z_{\lambda_n} - Z_{\lambda_{n-1}}$. It's often easier to control these chain links than to control $Z_\lambda - Z_{\lambda'}$ directly. This approach leads to two general bounds for empirical processes. On the one hand, there is the classical "Entropy Bound" (see for example [Tal05], [vdVW00] and references therein). Its integral version as stated in [vdVW00] is introduced and refined at the beginning of the second part. Then, we apply this bound to the problem stated above where we follow ideas given in [Car85] for some entropy calculations. On the other hand, there are "Majorizing Measures" (see for example [RT88], [Tal94] and [Tal96]). They are introduced and applied in the third part. Majorizing Measures are rather difficult to use, however, we show that for highly correlated covariables they can lead to substantially better results.

We conclude this section with some notation and the main results.

Notation: For a pseudometric space (S, d) with unit ball B we denote the covering numbers by $N(S, d, \epsilon)$, i.e., $N(S, d, \epsilon)$ is the number of translates of ϵB needed to cover S. The logarithm of the covering numbers (as a function of ϵ) is called entropy. We define similarly $D(S, d, \epsilon)$ as the maximal number of ϵ-separated points in S. Obviously, $N(S, d, \epsilon) \leq D(S, d, \epsilon) \leq N(S, d, \frac{\epsilon}{2})$. And finally, if the pseudometric is induced by a seminorm, we occasionally write $N(S, \|\cdot\|, \epsilon)$ or $D(S, \|\cdot\|, \epsilon)$.

We are mainly interested in the pseudometric space (Θ, d) with $d(\theta, \theta') := \|(\phi_\theta(x_1) - \phi_{\theta'}(x_1), ..., \phi_\theta(x_n) - \phi_{\theta'}(x_n))^T\|_2$, where $x := (x_1, ..., x_n) \in \mathfrak{X}^n$ for an arbitrary set $\mathfrak{X}$ and

$\{\phi_\theta : \mathfrak{X} \to \mathbb{R} : \theta \in \Theta\}$ is a set of functions and we define $X_\theta(x) := \sum_{i=1}^n \epsilon_i \phi_\theta(x_i)$ for simplicity. The choice for the pseudometric d is motivated by the fact that $\{X_\theta(x) : \theta \in \Theta\}$ is sub-Gaussian with respect to d due to Hoeffding's inequality, that is

$$\mathbb{P}(|X_\theta - X_{\theta'}| > u) \leq 2 \exp\left(-\frac{u^2}{2d(\theta, \theta')}\right) \ \forall \theta, \theta' \in \Theta.$$

In other words, the tail behavior is as for Gaussian processes.

Main Results: We derive upper bounds for the quantity (1) under three different sets of assumptions. We are not aware of equally sharp bounds in the literature.

In Section 2.2, we derive a bound under the assumption that $\{\phi_\theta : \mathfrak{X} \to \mathbb{R} : \theta \in \Theta\}$ has a certain contraction property:

Theorem 1.1. *If there exists a function $A : \mathfrak{X}^n \to \mathbb{R}$ fulfilling*

$$d(\theta, \theta') \leq \sqrt{n} A(x) \|\theta - \theta'\|_2 \ \forall \theta, \theta' \in \Theta \tag{2}$$

then there is a universal constant K such that for $\theta_0 \in \Theta$ arbitrary

$$\mathbb{E} \sup_{\theta \in \Theta} |X_\theta(x)| \leq \mathbb{E}|X_{\theta_0}(x)| + K \sqrt{n \log(l+1)} \log(n+1) A(x) M. \tag{3}$$

In the linear case, the $\log(n+1)$ in (3) can be omitted and the contraction property (2) can be relaxed. This is stated in the following theorem we prove in Section 2.3:

Theorem 1.2. *Let $\psi_j : \mathfrak{X} \to \mathbb{R}$ be arbitrary functions for $j = 1, ..., l$. If $\phi_\theta(x_i) = \sum_{j=1}^l \psi_j(x_i)\theta_j$ and if $A : \mathfrak{X}^n \to \mathbb{R}$ fulfills*

$$d(\theta, 0) \leq \sqrt{n} A(x) M \ \forall \theta \in \Theta$$

there is a universal constant K such that

$$\mathbb{E} \sup_{\theta \in \Theta} |X_\theta(x)| \leq K \sqrt{n \log(l+1)} A(x) M.$$

For strongly correlated covariables, we can improve on these bounds. We show this in Section 3.2 with the help of Majorizing Measures. To state the result, we let $X' \in \mathbb{R}^{n \times l'}$, $X'' \in \mathbb{R}^{n \times l''}$. Furthermore, we denote the i-th row of X' (X'' resp.) by x_i' (x_i'' resp.), the columns by y_i' (y_i'' resp.) and we set $\theta = (\theta', \theta'')$. We then impose the usual normalization on the matrices, that is $\|y_i'\|_2 = \|y_i''\|_2 = \sqrt{n}$ and state the following result:

Theorem 1.3. *Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a contraction w.r.t. the Euclidean metric. If there are orthogonal matrices R', R'' such that for all i*

$$\sum_{j=1}^n j \frac{(R' y_i')_j^2}{n}, \ \sum_{j=1}^n j \frac{(R'' y_i'')_j^2}{n} \leq 1 \tag{4}$$

then there is a universal constant K such that for $\theta_0 \in \Theta$ arbitrary

$$\mathbb{E} \sup_{\theta \in \Theta} \left| \sum_{i=1}^{n} \epsilon_i g((x_i')^T \theta', (x_i'')^T \theta'') \right| \leq \mathbb{E} \left| \sum_{i=1}^{n} \epsilon_i g((x_i')^T \theta_0', (x_i'')^T \theta_0'') \right| + K \sqrt{n \log(n+1)} M.$$

So, the factor $\sqrt{n \log(l+1)} \log(n+1)$ in the bound (3) can be replaced by $\sqrt{n \log(n+1)}$ in this case. The required correlation is expressed by assumption (4): It means, that the columns of the matrices X' and X'' can be enveloped by small ellipsoids. The matrices R' and R'' are the transformations that bring these ellipsoids on the standard form.

2 Entropy Bounds

In this part, we introduce entropy bounds and apply them to Rademacher processes. In the first section, we prove adapted versions of two classical entropy results. The second and the third sections are devoted to the proofs of Theorem 1.1 and Theorem 1.2 and a simple example.

2.1 Refinement of Entropy Bounds

Here, we introduce slightly modified versions of two classical entropy bounds for empirical processes (see e.g. [vdVW00] Theorem 2.2.4 and Corollary 2.2.8). The modification is the lower bound for the integration. For convenience, we give the proofs in detail, although they follow closely the ones given in [vdVW00].

Beforehand, we recall the definition of the Orlicz norm $\|X\|_\Psi$ for a non-decreasing and convex function Ψ with $\Psi(0) = 0$:

$$\|X\|_\Psi := \inf\{A > 0 : \mathbb{E}\Psi\left(\frac{|X|}{A}\right) \leq 1\}.$$

We are then able to formulate and prove an important entropy bound:

Lemma 2.1. *Let $\Psi : \mathbb{R} \to \mathbb{R}$ be a convex, non-decreasing and non-constant function with $\Psi(0) = 0$ and*

$$\limsup_{x,y \to \infty} \frac{\Psi(x)\Psi(y)}{\Psi(cxy)} < \infty$$

for a constant c. Define $\Psi(\infty) := \infty$, $\Psi^{-1}(y) := \sup\{x : \Psi(x) \leq y\}$ and assume $\Psi^{-1}(1) > 0$. Furthermore, let $\{X_t : t \in T\}$ be a stochastic process with

$$\|X_s - X_t\|_\Psi \leq C d(s,t) \ \forall s,t \in T$$

and

$$|X_s - X_t| \leq \alpha d(s,t) \ \forall s,t \in T \tag{5}$$

for a pseudometric d and positiv constants C and α. Then there are universal functions $K \equiv K(C, \Psi)$ and $U \equiv U(C, \Psi, \alpha)$ such that for all $0 < \eta \leq \delta$

$$\| \sup_{d(s,t)\leq\delta} |X_s - X_t| \|_\Psi \leq K \left(\int_{\frac{\eta}{U}}^{\frac{\eta}{2}} \Psi^{-1}(D(T,d,\epsilon))d\epsilon + \delta\Psi^{-1}(D^2(T,d,\eta)) \right). \tag{6}$$

Comparing this to [vdVW00], note that we introduced the additional condition (5). This is to establish the lower integral bound in the inequality (6).

Proof. We may assume that the covering numbers for $\epsilon > \frac{\eta}{U}$ and the corresponding integral in (6) are finite since the inequality is trivial otherwise. We then fix $\eta \in \mathbb{R}^+$ and $k \in \mathbb{N}$ and construct nested sets $T_0 \subset T_1 \subset ... \subset T_{k+1} \subset T$ such that for every $j \leq k+1$ T_j is maximal w.r.t. $d(s,t) > \eta 2^{-j}$ for all $s, t \in T_j$.
According to the definition of covering numbers, it holds that $|T_j| \leq D(T, d, \eta 2^{-j})$. We will assume $U 2^{-(k+1)} > 1$ (U will be defined later) and hence finitely many elements in every set, this will be justified later. Now, we will assign each point $t_{j+1} \in T_{j+1}$ to a unique point $t_j \in T_j$ such that $d(t_{j+1}, t_j) \leq \eta 2^{-j}$. In this way, we define for all $t_{k+1} \in T_{k+1}$ chains $t_{k+1} \mapsto ... \mapsto t_0 \in T_0$ and use the notation $c(t_{k+1}) := \{t_{k+1}, ..., t_0\}$.
Let $s_{k+1}, t_{k+1} \in T_{k+1}$. We then get for elements of these chains

$$\begin{aligned}
|(X_{s_{k+1}} - X_{s_0}) - (X_{t_{k+1}} - X_{t_0})| &= |\sum_{j=0}^{k}(X_{s_{j+1}} - X_{s_j}) - \sum_{j=0}^{k}(X_{t_{j+1}} - X_{t_j})| \\
&\leq \sum_{j=0}^{k} |X_{s_{j+1}} - X_{s_j}| + \sum_{j=0}^{k} |X_{t_{j+1}} - X_{t_j}| \\
&\leq 2\sum_{j=0}^{k} \max\{|X_u - X_v| : u \in T_{j+1}, v \in T_j \cap c(u)\}.
\end{aligned}$$

Applying Lemma 2.2.2 of [vdVW00], we find a constant K depending on Ψ only such that

$$\begin{aligned}
&\| \max |(X_{s_{k+1}} - X_{s_0}) - (X_{t_{k+1}} - X_{t_0})| \|_\Psi \\
\leq& 2\sum_{j=0}^{k} \| \max\{|X_u - X_v| : u \in T_{j+1}, v \in T_j \cap c(u)\} \|_\Psi \\
\leq& 2K\sum_{j=0}^{k} \Psi^{-1}(|T_{j+1}|) \max\{\|X_u - X_v\|_\Psi : u \in T_{j+1}, v \in T_j \cap c(u)\} \\
\leq& 2KC\sum_{j=1}^{k+1} \Psi^{-1}(D(T,d,\eta 2^{-j}))\eta 2^{-j+1} \\
\leq& 8KC \int_{\eta 2^{-(k+2)}}^{\frac{\eta}{2}} \Psi^{-1}(D(T,d,\epsilon))d\epsilon.
\end{aligned}$$

In the first line, the maximum is taken over all $s_{k+1}, t_{k+1} \in T_{k+1}$ and their associated points in T_0. We then note that for $\delta \geq \eta$

$$\begin{aligned}
&\| \max\{|X_s - X_t| : s,t \in T_{k+1} : d(s,t) \leq \delta\}\|_\Psi \\
\leq &\| \max\{|(X_s - X_{s_0}) - (X_t - X_{t_0})| : s,t \in T_{k+1} : d(s,t) \leq \delta\}\|_\Psi \\
+ &\| \max\{|X_{s_0} - X_{t_0}| : s_0, t_0 \in T_0, s,t \in T_{k+1}, s_0 \in c(s), t_0 \in c(t)\}\|_\Psi.
\end{aligned}$$

The first term on the r.h.s. of the last display is bounded according to what we have done above. The second term may be rewritten using

$$|X_{s_0} - X_{t_0}| \leq |(X_{s_0} - X_{s_{k+1}}) - (X_{t_0} - X_{t_{k+1}})| + |X_{s_{k+1}} - X_{t_{k+1}}|.$$

Here, we assign to each $s_0 \in T_0$ and each $t_0 \in T_0$ a fixed $s_{k+1} \in T_{k+1}$, $t_{k+1} \in T_{k+1}$ respectively, such that $s_0 \in c(s)$ and $t_0 \in c(t)$. We demand furthermore, that $d(s_{k+1}, t_{k+1}) \leq \delta$. This yields together with Lemma 2.2.2 of [vdVW00]

$$\begin{aligned}
&\| \max\{|X_s - X_t| : s,t \in T_{k+1}, d(s,t) \leq \delta\}\|_\Psi \\
\leq &16KC \int_{\eta 2^{-(k+2)}}^{\frac{\eta}{2}} \Psi^{-1}(D(T,d,\epsilon))d\epsilon + \| \max |X_{s_{k+1}} - X_{t_{k+1}}|\|_\Psi \\
\leq &16KC \int_{\eta 2^{-(k+2)}}^{\frac{\eta}{2}} \Psi^{-1}(D(T,d,\epsilon))d\epsilon + K\Psi^{-1}(D^2(T,d,\eta)) \max \|X_{s_{k+1}} - X_{t_{k+1}}\|_\Psi \\
\leq &16KC \int_{\eta 2^{-(k+2)}}^{\frac{\eta}{2}} \Psi^{-1}(D(T,d,\epsilon))d\epsilon + KC\delta\Psi^{-1}(D^2(T,d,\eta)).
\end{aligned}$$

The maximum in the second line is taken as described above. We then note that

$$\begin{aligned}
\| \sup_{d(s,t)\leq\delta} |X_s - X_t|\|_\Psi = &\| \sup_{d(s,t)\leq\delta} |(X_s - X_{s^*}) - (X_t - X_{t^*}) + (X_{s^*} - X_{t^*})|\|_\Psi \\
\leq &2\| \sup_{s\in T} |X_s - X_{s^*}|\|_\Psi + \| \max\{|X_s - X_t| : s,t \in T_{k+1}, d(s,t) \leq 3\delta\}\|_\Psi
\end{aligned}$$

where we define $s^* := \arg\min_{s' \in T_{k+1}} d(s', s)$ and $t^* := \arg\min_{t' \in T_{k+1}} d(t', t)$ and use

$$d(s^*, t^*) \leq d(s^*, s) + d(s,t) + d(t, t^*) \leq 3\delta.$$

We find moreover

$$\begin{aligned}
\| \sup_{s\in T} |X_s - X_{s^*}|\|_\Psi &= \inf\{A > 0 : \mathbb{E}\Psi(\sup_{s\in T} |X_s - X_{s^*}|/A) \leq 1\} \\
&\leq \frac{\alpha\eta 2^{-(k+1)}}{\Psi^{-1}(1)}.
\end{aligned}$$

We may assume w.l.o.g. that T is not empty and $C, K > 0$. So there is a $k_0 \in \mathbb{N}$

(depending only on Ψ and α) such that $\frac{\alpha 2^{-(k_0+1)}}{\Psi^{-1}(1)} \leq \frac{K}{4}\Psi^{-1}(1)$. Then,

$$\begin{aligned}\frac{\alpha\eta 2^{-(k_0+1)}C}{\Psi^{-1}(1)} &\leq KC\frac{\eta}{4}\Psi^{-1}(1)\\ &\leq KC\frac{\eta}{2}(1-2^{-(k_0+1)})\Psi^{-1}(1)\\ &\leq KC\int_{\eta 2^{-(k_0+2)}}^{\frac{\eta}{2}}\Psi^{-1}(D(T,d,\epsilon))d\epsilon.\end{aligned}$$

We define $U := 2^{k_0+2}$ to conclude the proof. □

Because we often do not need the generality of Lemma 2.1, we derive in the following a result for the important special case of sub-Gaussian processes:

Lemma 2.2. *Let $\{X_t : t \in T\}$ be a sub-Gaussian process w.r.t. a pseudometric d such that*

$$|X_s - X_t| \leq \alpha d(s,t)\ \forall s,t \in T$$

for a constant α. Then there exists a function $U \equiv U(\alpha)$ and a universal constant K such that for all $\delta > 0$ and $t_0 \in T$ arbitrary

$$\mathbb{E}\sup_{t:d(t,t_0)\leq\delta}|X_t| \leq \mathbb{E}|X_{t_0}| + K\int_{\frac{\delta}{U}}^{\frac{\delta}{2}}\sqrt{\log(1+D(T,d,\epsilon))}d\epsilon. \tag{7}$$

Proof. We apply Lemma 2.1 to $\Psi(x) := e^{x^2}-1$. The function Ψ is convex and increasing and $\Psi(0)=0$. It holds that

$$\limsup_{x,y\to\infty}\frac{\Psi(x)\Psi(y)}{\Psi(xy)} < \infty$$

and

$$\|X_s - X_t\|_\Psi \leq \sqrt{6}d(s,t)\ \forall s,t \in T.$$

So, the conditions of Lemma 2.1 are met. We then set $\eta = \delta$ in Lemma 2.1 and note that

$$\Psi^{-1}(m^2) = \sqrt{\log(1+m^2)} \leq \sqrt{\log(1+m)^2} = \sqrt{2}\Psi^{-1}(m).$$

So there is a universal constant K' such that (recall that $U \geq 4$, cf. proof of Lemma 2.1)

$$\|\sup_{s,t:d(s,t)\leq\delta}|X_s - X_t|\|_\Psi \leq K'\int_{\frac{\delta}{U}}^{\frac{\delta}{2}}\sqrt{\log(1+D(T,d,\epsilon))}d\epsilon.$$

Since $\sqrt{\log 2}\cdot\mathbb{E}|X| \leq \|X\|_\Psi$ for any random variable X, there is a constant K such that

$$\mathbb{E}\sup_{s,t:d(s,t)\leq\delta}|X_s - X_t| \leq K\int_{\frac{\delta}{U}}^{\frac{\delta}{2}}\sqrt{\log(1+D(T,d,\epsilon))}d\epsilon.$$

We conclude the proof by noting that for any t_0

$$\mathbb{E}\sup_{t:d(t,t_0)\leq\delta}|X_t| - \mathbb{E}|X_{t_0}| \leq \mathbb{E}\sup_{s,t:d(s,t)\leq\delta}|X_s - X_t|.$$

□

2.2 Proof of Theorem 1.1

The proof of Theorem 1.1 has two main ingredients: First, the entropy bound of Lemma 2.1 and second, some subtle entropy estimates. For the entropy estimates, we rely on ideas given in Lemma 1 of [Car85].

Proof of Theorem 1.1. To simplify the notation, we set $X_\theta := X_\theta(x)$ and $A := A(x)$. We then note that, as a consequence of Hoeffding's inequality, $\{X_\theta : \theta \in \Theta\}$ is sub-Gaussian with respect to the pseudometric

$$d(\theta, \theta') := \|(\phi_\theta(x_1) - \phi_{\theta'}(x_1), ..., \phi_\theta(x_n) - \phi_{\theta'}(x_n))^T\|_2.$$

We find that

$$|X_\theta - X_{\theta'}| \leq \sqrt{n} d(\theta, \theta') \leq nA\|\theta - \theta'\|_2 \ \forall \theta, \theta' \in \Theta. \tag{8}$$

Now, we want to calculate the entropy linked with the stochastic process and the pseudometric d. To this end, we define

$$V := \{e_1, ..., e_{2l}\} \subset \mathbb{R}^l$$

using the notation $(e_i)_j := \delta_{ij}$ for $i \leq l$, where δ_{ij} is the Kronecker symbol, and $e_i := -e_{2l-i+1}$ for $i > l$. So Θ is the set $\{\theta \in \mathbb{R}^l : \exists \lambda \in \mathbb{R}^{2l}, \|\lambda\|_1 \leq M, \theta = \sum_{i=1}^{2l} \lambda_i e_i\}$. We then fix a $\lambda \in \mathbb{R}^{2l}$ such that $\|\lambda\|_1 \leq M$. Define independent random variables $Y_1, ..., Y_k \in V \cup \vec{0}$ with (following [Car85])

$$\mathbb{P}(Y_i = e_j) = \frac{|\lambda_j|}{M} \ \forall i = 1, ..., k, j = 1, ..., 2l$$

and

$$\mathbb{P}(Y_i = \vec{0}) = 1 - \sum_{j=1}^{2l} \frac{|\lambda_j|}{M}.$$

We obtain

$$\mathbb{E}Y_i = \frac{1}{M} \sum_{j=1}^{2l} |\lambda_j| e_j \in \Theta \ \forall i.$$

Next, we set $\overline{Y}_k := \frac{1}{k} \sum_{i=1}^k Y_i \in \Theta$. One may check that

$$\mathbb{E}[d(M\overline{Y}_k, M\mathbb{E}Y_1)^2] \leq \frac{4nA^2M^2}{k}$$

using the contraction property (2). So, the distance of at least one realization of $M\overline{Y}_k$ to $M\mathbb{E}Y_1$ is smaller or equal to $2\sqrt{\frac{n}{k}}AM$. For the (at most $\binom{2l+k-1}{k}$) realizations of $M\overline{Y}_k$ and $M\mathbb{E}Y_1$ it holds that $\forall \theta \in \Theta \ \exists \lambda : \|\lambda\|_1 \leq M, \theta = M \sum_{j=1}^{2l} \frac{|\lambda_j|}{M} e_j$. Hence, using Stirling's inequalities, we get

$$N\left(\Theta, d, 2\sqrt{\frac{n}{k}}AM\right) \leq \binom{2l + k - 1}{k} \leq \left(e + \frac{2el}{k}\right)^k.$$

Therefore,

$$N(\Theta, d, \epsilon) \leq \left(e + \frac{el\epsilon^2}{2nM^2A^2} \right)^{\frac{4nM^2A^2}{\epsilon^2}+1}$$

when we choose $k := \lceil \frac{4nM^2A^2}{\epsilon^2} \rceil$. Consequently,

$$D(\Theta, d, \epsilon) \leq \left(e + \frac{el\epsilon^2}{8nM^2A^2} \right)^{\frac{16nM^2A^2}{\epsilon^2}+1}.$$

We may now use Lemma 2.1 and get for a universal constant K and a constant U depending only on $\sqrt{n}$ (see condition (5) and inequality (8))

$$\mathbb{E} \sup_{\theta \in \Theta} |X_\theta| - \mathbb{E}|X_{\theta_0}| \leq K \int_{\frac{\sqrt{n}AM}{U}}^{\sqrt{n}AM} \sqrt{\log\left(1 + D(\Theta, d, \epsilon)\right)} d\epsilon.$$

Regarding the last part of the proof of Lemma 2.1 we find a universal constant V such that $U = \sqrt{n}V$. The results then follows by a simple calculation. □

2.3 Proof of Theorem 1.2 and an Example

In the linear case, we can get rid of one of the logarithms. This is because we can transform the parameter space into a lower dimensional one. We note that in the proof of this lemma, the lower bounds for the integrals in Lemma 2.1 and Lemma 2.2 are not necessary. Additionally, no difficult entropy estimates have to be made.

Proof of Theorem 1.2. Again, we set $X_\theta := X_\theta(x)$ and $A := A(x)$ and note that

$$\sup_{\theta \in \Theta} |X_\theta| = \sup_{\theta \in \Theta} |\theta^T a|$$

with $a := (\sum_{i=1}^n \epsilon_i \psi_1(x_i), ..., \sum_{i=1}^n \epsilon_i \psi_l(x_i))^T \in \mathbb{R}^l$. The map $\theta \to |\theta^T a|$ attains its maximum on Θ at θ_0 where $(\theta_0)_i := M\delta_{ip}$ with p such that $|a_p| \geq |a_m|$ for all $m = 1, ..., l$. So we have

$$\mathbb{E} \sup_{\theta \in \Theta} |X_\theta| = \mathbb{E} \sup_{\theta \in \Theta'} |X_\theta|$$

for $\Theta' := \{(M, 0, ..., 0)^T, ..., (0, ..., 0, M)^T, (0, ..., 0)^T\}$. As a consequence of Hoeffding's inequality, $\{X_\theta : \theta \in \Theta'\}$ is sub-Gaussian with respect to the pseudometric $d(\theta, \theta') := \|(\phi_\theta(x_1) - \phi_{\theta'}(x_1), ..., \phi_\theta(x_n) - \phi_{\theta'}(x_n))^T\|_2$ and it holds for all θ, θ' that $d(\theta, 0) \leq \sqrt{n}M$. Hence, according to Lemma 2.1, we get for a universal constant K

$$\mathbb{E} \sup_{\theta \in \Theta} |X_\theta| \leq K \int_0^{\sqrt{n}AM} \sqrt{\log\left(1 + D(\Theta', d, \epsilon)\right)} d\epsilon.$$

The result follows then using $D(\Theta', d, \epsilon) \leq |\Theta'| = l + 1$. □

Finally, we give a simple application:

Example 2.1. *Let $X \in \mathbb{R}^{n\times l}$ be normalized such that the columns have Euclidean norm $\sqrt{n}$. Moreover, define $\vec{\epsilon} := (\epsilon_1, ..., \epsilon_n)$ with Rademacher variables ϵ_i. Then, for $X_\theta := \vec{\epsilon}^{\,T} X\theta,\ \theta \in \Theta = \{\theta \in \mathbb{R}^l : \|\theta\|_1 \leq M$, there is a universal constant K such that*

$$\mathbb{E} \sup_{\theta\in\Theta} |X_\theta| \leq K\sqrt{n \log (l+1)} M.$$

3 The Majorizing Measures Bound

In this part, we recall the Majorizing Measures Bound and some consequence such as the Ellipsoid Theorem. We then apply these tools to prove Theorem 1.3.

3.1 Majorizing Measures

Majorizing Measures are known to work well in situations where we have unit balls of p-convex Banach spaces as index sets (see [GMPTJ08] for an example and [Pis89] or [LT79] for the definitions of p-convexity, p-type and related terms). Here, we recall the most important bounds arising in this scope. For the proofs and more detailed introductions we refer to [RT88], [Tal94] and [Tal96].

We begin with a basic definition:

Definition 3.1. *Let $(T,\bar{d})$ be a metric space and $\beta > 0$. We set*

$$\gamma_\beta(T, \bar{d}) := \inf \left\{ \sup_{t\in T} \left(\int_0^\infty \epsilon^{\beta-1} \left(\log \frac{1}{\mu(B(\bar{d}, t, \epsilon))} \right)^{\frac{\beta}{2}} d\epsilon \right)^{\frac{1}{\beta}} \right\},$$

where $B(\bar{d}, t, \epsilon)$ is the ball w.r.t. $\bar{d}$ around t with radius ϵ and the infimum is taken over all probability measures μ on the Borel-σ-algebra of T.

We then recall the following bounds:

Lemma 3.1. (The Majorizing Measures Bound) *Any sub-Gaussian process fulfills*

$$\mathbb{E} \sup_{t\in T} X_t \leq K\gamma_1(T, \bar{d})$$

for a universal constant K.

Lemma 3.2. (The Ellipsoid Theorem) *Let the metric $\bar{d}$ be induced by the norm on $l^2(\mathbb{N})$. Then, for*

$$E := \{(t_i)_{i\geq 1} : \sum_{i\geq 1} \frac{t_i^2}{a_i^2} \leq 1\} \subset l^2$$

with $(a_i)_{i\geq 1} \in l^2(\mathbb{N})$ positive and non-increasing we have

$$\gamma_2(E, \bar{d}) \leq K \sup_{i\geq 1} a_i\sqrt{i} \tag{9}$$

for a universal constant K.

Using Hölders inequality, the bound (9) may be used to give an upper bound for $\gamma_1(T, \bar{d})$. Finally, it holds that

Lemma 3.3. *Consider a metric space (T,$\bar{d}$) and a subset S of T. Then,*

$$\gamma_\beta(S, \bar{d}) \leq 2\gamma_\beta(T, \bar{d}).$$

3.2 Proof of Theorem 1.3

Now, we show how the process of Theorem 1.3 can be rewritten such that the relevant set is an ellipsoid and how the bounds stated above can then be applied. To find reasonable results, however, we have to assume strong correlation among the covariables. By this, we mean that the columns of the corresponding matrices are not too different. Or, more precisely, that the columns regarded as vectors can be collectively enveloped by a small ellipsoid.

At first, we state a well known fact:

Proposition 3.1. *Let $\{X_t : t \in T\}$ be a stochastic process with an arbitrary index set T. Assume that the $\mathbb{E}\sup_{t\in T} X_t = \mathbb{E}\sup_{t\in T}(-X_t)$. Then,*

$$\mathbb{E}\sup_{t\in T}|X_t| - \mathbb{E}|X_{t_0}| \leq 2\mathbb{E}\sup_{t\in T} X_t$$

for $t_0 \in T$ arbitrary.

Moreover, we set $\frac{0}{0} := 0$ and we denote by sconv A the symmetric convex hull of a set A. We are then prepared to give the proof of the theorem:

Proof of Theorem 1.3. Setting

$$T' := M \cdot \text{sconv}\,\{y'_1, ..., y'_{l'}\}$$
$$T'' := M \cdot \text{sconv}\,\{y''_1, ..., y''_{l''}\}$$

we obtain

$$\mathbb{E}\sup_{\theta\in\Theta}\left|\sum_{i=1}^n \epsilon_i g((x'_i)^T\theta', (x''_i)^T\theta'')\right| \leq \mathbb{E}\sup_{t\in T'\times T''}\left|\sum_{i=1}^n \epsilon_i g(t'_i, t''_i)\right|.$$

Next we define $a_i^2 := \frac{4n}{i}\cdot M^2$, $(\Pi'(t))_i := t_{2i-1}$ and $(\Pi''(t))_i := t_{2i}$. Furthermore,

$$E := \{t \in \mathbb{R}^{2n} : \sum_{i=1}^{2n}\frac{t_i^2}{a_i^2} \leq 1\}.$$

Then,

$$\mathbb{E}\sup_{\theta\in\Theta}\left|\sum_{i=1}^n \epsilon_i g((x'_i)^T\theta', (x''_i)^T\theta'')\right| \leq \mathbb{E}\sup_{t\in E}\left|\sum_{i=1}^n \epsilon_i g((R'^{-1}\Pi'(t))_i, (R''^{-1}\Pi''(t))_i)\right|.$$

To simplify the notation, we define

$$g_i(t) := g((R'^{-1}\Pi'(t))_i, (R''^{-1}\Pi''(t))_i)$$

and we note that since g is a contraction

$$\bar{d}(t,\tilde{t}) := \sqrt{\sum_{i=1}^{n}(g_i(t)-g_i(\tilde{t}))^2} \leq \|t-\tilde{t}\|_2 =: d_2(t,\tilde{t}). \tag{10}$$

Now, let S be a maximal subset of E such that $\bar{d}(t,\tilde{t}) > M$ for all $t,\tilde{t} \in S, t \neq \tilde{t}$. Consequently, (S,d_2) is a metric space and we have due to Cauchy-Schwarz' inequality

$$\sup_{t\in E}\left|\sum_{i=1}^{n}\epsilon_i g_i(t)\right| \leq \sqrt{n}M + \sup_{t\in S}\left|\sum_{i=1}^{n}\epsilon_i g_i(t)\right|.$$

With regard to Proposition 3.1, the quantity to calculate is

$$\mathbb{E}\sup_{t\in S}\sum_{i=1}^{n}\epsilon_i g_i(t).$$

To bound this quantity, we apply Hoeffding's inequality, the contraction property (10) and Lemma 3.1 to obtain for a universal constant K

$$\mathbb{E}\sup_{t\in S}\sum_{i=1}^{n}\epsilon_i g_i(t) \leq K\gamma_1(S,d_2).$$

Moreover, $d_2^2(t,0) \leq \sum_{i=1}^{2n} a_i^2 \leq 8n^2M^2$, so that we arrive at (using Hölders inequality)

$$\begin{aligned}&\int_0^\infty \sqrt{\log\frac{1}{\mu(B(d_2,t,\epsilon))}}d\epsilon \\ \leq &\int_0^M \sqrt{\log\frac{1}{\mu(B(d_2,t,\epsilon))}}d\epsilon + \left(\int_M^{4nM}\frac{d\epsilon}{\epsilon}\right)^{\frac{1}{2}}\left(\int_0^\infty \epsilon\log\frac{1}{\mu(B(d_2,t,\epsilon))}d\epsilon\right)^{\frac{1}{2}}\end{aligned}$$

We stress, that the balls are with respect to the set S. Finally,

$$\sqrt{2}\left(\int_0^\infty \epsilon\log\frac{1}{\mu(B(d_2,t,\epsilon))}d\epsilon\right)^{\frac{1}{2}} \geq \int_0^M \sqrt{\log\frac{1}{\mu(B(d_2,t,\epsilon))}}d\epsilon.$$

Thus, the proof can be concluded using Lemma 3.2 and Lemma 3.3. □

4 Conclusion

Classical entropy bounds have proved to be a simple and useful tool in many applications. However, Majorizing Measures are a priori more powerful in the treatment of empirical processes. They are known to outmatch the classical entropy bounds for unit balls of p-convex Banach spaces as index sets. While this is true, the unit ball of $(\mathbb{R}^l, \|\cdot\|_1)$ is not p-convex. So far, we only found reasonable results with Majorizing Measures by invoking high correlation. The results were in this case independent of the dimension l, which is quite important since we often assume $l \gg n$.

Acknowledgments

I thank Sara van de Geer for the excellent support. Furthermore, I thank Mohamed Hebiri for his interest and some helpful suggestions.

References

[Car85] B. Carl. Inequalities of bernstein-jackson-type and the degree of compactness of operators in banach spaces. *Annales de l'institut Fourier*, 35, no.3, 1985.

[GMPTJ08] O. Guedon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Revista Matematica Iberoamericana*, 24, no. 3, 2008.

[LT79] J. Lindenstrauss and L. Tzfariri. *Classical Banach Spaces II.* Springer, 1979. ISBN 3-540-08888-1.

[Pis89] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press, 1989. ISBN 0-521-364655.

[RT88] W.T. Rhee and M. Talagrand. Exact bounds for the stochastic upward matching problem. *Transactions of the American Mathematical Society*, May 1988.

[Tal94] M. Talagrand. Matching theorems and empirical discrepancy computations using majorizing measures. *Journal of the American Mathematical Society*, April 1994.

[Tal96] M. Talagrand. Majorizing measures: The generic chaining. *Annals of Probability*, July 1996.

[Tal05] M. Talagrand. *The Generic Chaining.* Springer, 2005. ISBN 3-540-24518-9.

[vdVW00] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, 2000. ISBN 0-387-94640-3.